Simultaneous Estimation of Multinomial Cell Probabilities

by Stephen E. Fienberg and Paul W. Holland*

Technical Report No. 188

*Stephen E. Fienberg is Associate Professor and Chairman, Department of Applied Statistics, University of Minnesota, St. Paul, Minn. 55101. Paul W. Holland is a Senior Research Associate, N.B.E.R. Computer Research Center, Cambridge, Mass. 02139, and Lecturer on Statistics, Harvard University, Cambridge, Mass. 02138. This research was supported in part by a grant from the Alfred P. Sloan Foundation to the Department of Theoretical Biology, University of Chicago; by Research Grant No. NSF-GP-16071 from the National Science Foundation to the Department of Statistics, University of Chicago; by Research Grant No. NSF-GS-2044x1 to the Department of Statistics, Harvard University; and a Faculty Research Grant from the Social Science Research Council. The authors wish to thank Robert Fay for computational assistance, Michael Sutherland for technical comments, and R.R. Bahadur, Arthur Dempster, Frederick Mosteller, and Herbert Weisberg for comments on earlier versions of this work.

Abstract

We propose a new estimator, $\underset{\sim}{p}^*$, of the multinomial parameter vector, and show that it is a better choice in most situations than the usual estimator, $\hat{\underset{\sim}{p}}$ (the vector of observed proportions). Using squared-error loss, we examine the risk functions (expected loss) of these estimators in three ways: (a) we compare the exact risk functions for selected small sample sizes; (b) we show the equivalence of $\underset{\sim}{p}^*$ and $\hat{\underset{\sim}{p}}$ when the sample size tends towards infinity and the number of cells remains constant; (c) we approximate the risk functions in a novel asymptotic framework in which the number of cells is large and the number of observations per cell is moderate. These approximations reveal the general superiority of $\underset{\sim}{p}^*$ over $\hat{\underset{\sim}{p}}$ in large sparse multinomials. This novel asymptotic framework is also of interest in its own right, and may provide insight in other multinomial problems.

1. Introduction

In the analysis of multinomial data, we frequently wish to provide a table of smoothed cell frequencies that can be used for other purposes, such as creating standardized rates, or that are more suitable for transformation to the various linearizing scales such as logarithms, logits or probits. For many of these purposes, the presence of zero observed counts in some of the cells creates serious obstacles.

In the analysis of binomial data, much attention has been focused on devices which eliminate the problematic zero counts. Gart and Zweifel [1967] discuss adjustments to the observed cell counts for use in estimates of the logit, log (p/q), where p is the binomial probability. These adjustments are of special interest when the observed proportions, $\hat{p}$ and $\hat{q}$, are either 0 or 1, and log $(\hat{p}/\hat{q})$ is either $-\infty$ or $+\infty$. They typically consist of adding a constant (such as $\frac{1}{2}$ or 1) to each observed count, or to each zero observed count. Freeman and Tukey [1950] propose yet a different approach for use with the arc-sine transformation for variance stabilization.

In the analysis of multinomially distributed, cross-classified data, zero observed counts create problems in the study of loglinear models. Goodman [1970,1971] suggests the use of techniques analogous to those for logit analysis in order to replace the logarithms and inverses of observed counts for parameter and variance estimation. He suggests adding $\frac{1}{2}$ count to every cell. In an earlier paper, Goodman [1964] suggested adding a count of 1 to each cell for estimating sums of reciprocals of cell probabilities. Grizzle, Starmer and Koch [1969] and Johnson and Koch [1971] encounter a different, yet related, problem with zero counts. Their analysis of loglinear models relies upon a weighted least squares approach which requires non-zero counts. These authors replace zero counts

by 1/t where t is the total number of cells. Both the Goodman and Grizzle-Starmer-Koch suggestions are now widely used in practice.

The above discussion suggests the need for a general purpose method for smoothing tables of observed counts, and thus for eliminating problematic zeros. In this paper we discuss such a method, which we view in terms of the simultaneous estimation of all the multinomial cell probabilities. Our method leads to a class of possible multinomial estimators, and we discuss the properties of some of these estimators.

In the next two sections, we introduce our class of smoothed multinomial estimators using a heuristic geometrical argument and a two-stage Bayesian argument.

The remaining sections of this paper are organized as follows. Section 4 gives two kinds of asymptotic results that help clarify the type and extent of the improvement in the estimation of p that is possible by use of the estimators developed in Section 2. In Section 5 we give some exact comparisons of the risk functions in small samples and small dimensions. Section 6 considers some examples in which we apply a differential smoothing technique that is not constant from sample to sample but depends on the sample itself.

A striking conclusion of the asymptotic analyses of Section 4 is that the method we propose is often superior to the common practice of adding $\frac{1}{2}$ to the observed count in each cell of a large, sparse table.

2. A Class of Smoothed Estimators

Let $\underset{\sim}{X} = (X_1,\ldots,X_t)$ have the multinomial distribution $M(n,\underset{\sim}{p})$ where $n = \sum_{i=1}^{t} X_i$ and $\underset{\sim}{p} = (p_1,\ldots,p_t)$ is the underlying vector of cell probabilities. The vector $\underset{\sim}{p}$ takes values in the parameter space, S_t, where

$$S_t = \{\underset{\sim}{p} = (p_1,\ldots,p_t): \quad p_i \geq 0 \text{ and } \sum_{i=1}^{t} p_i = 1\}$$

Let $\underset{\sim}{c} = (t^{-1},t^{-1},\ldots,t^{-1})$ denote the "center" of S_t.

We are concerned with the problem of the simultaneous estimation of all the components of $\underset{\sim}{p}$. Let $\underset{\sim}{T} = (T_1,\ldots,T_t)$ denote an estimator of the vector parameter, $\underset{\sim}{p}$. We adopt the squared distance from $\underset{\sim}{T}$ to $\underset{\sim}{p}$ as our loss function and we judge the ability of $\underset{\sim}{T}$ to estimate $\underset{\sim}{p}$ by the expected value of the squared distance, i.e. we use the risk function given by

$$R(\underset{\sim}{T},\underset{\sim}{p}) = nE\|\underset{\sim}{T} - \underset{\sim}{p}\|^2. \tag{2.1}$$

Other choices of the risk function $R(\underset{\sim}{T},\underset{\sim}{p})$ might be appropriate for some problems. A more general risk function of which (2.1) is a special case is

$$E\left[\sum_{i=1}^{t} c_i(T_i - p_i)^2\right] \tag{2.2}$$

where the c_i's may be constants or they may depend on $\underset{\sim}{T}$, on $\underset{\sim}{p}$, or on both. We have chosen (2.1) as our risk function in part for simplicity and by symmetry considerations, and in part for mathematical convenience. This risk function has also been used in other multinomial studies and in studies on simultaneous estimation for the mean of the multivariate normal distribution (see Efron and Morris, 1971 and 1972). It is important to note that the techniques suggested in this paper, both for generating estimators and for evaluating their properties, can also be used for other risk functions such as (2.2).

The usual estimator of $\underset{\sim}{p}$ is the vector of cell proportions

$$\hat{\underset{\sim}{p}} = n^{-1}\underset{\sim}{X}. \tag{2.3}$$

The risk function of $\hat{\underset{\sim}{p}}$ is easily shown to be

$$R(\hat{\underset{\sim}{p}},\underset{\sim}{p}) = nE\|\hat{\underset{\sim}{p}} - \underset{\sim}{p}\|^2 = (1 - \|\underset{\sim}{p}\|^2). \tag{2.4}$$

Each $\hat{p}_i$ is well-known to be the unique minimum variance unbiased estimate of p_i. From this fact it follows that if $\underset{\sim}{T}$ is any unbiased estimator of $\underset{\sim}{p}$ (i.e. $E(\underset{\sim}{T}) = \underset{\sim}{p}$) then the risk function of $\underset{\sim}{T}$ is never smaller than the risk function of $\hat{\underset{\sim}{p}}$ for any $\underset{\sim}{p} \in S_t$. Thus, no improvement in the estimation of $\underset{\sim}{p}$ can be achieved unless we leave the class of unbiased estimators. Furthermore, Johnson [1971] has shown that $\hat{\underset{\sim}{p}}$ is an admissible estimator of $\underset{\sim}{p}$ with respect to the risk function (2.1) so that there exists no biased estimator whose risk function is uniformly smaller than that of $\hat{\underset{\sim}{p}}$. Nevertheless, as Johnson points out, the reason $\hat{\underset{\sim}{p}}$ is admissible is not because it has small risk everywhere; rather it is due to the smallness of its risk on the boundary of the parameter space. The risk of $\hat{\underset{\sim}{p}}$ is smallest when $\underset{\sim}{p}$ has one component near unity. Hence one would expect to be able to improve upon $\hat{\underset{\sim}{p}}$ for those values of $\underset{\sim}{p}$ that are not so extreme. We shall show here that there exist estimators which provide a substantial improvement over $\hat{\underset{\sim}{p}}$ away from the boundary, and that the region of improvement in the parameter space increases as t becomes large.

To motivate our approach we make use of a heuristic geometrical argument which we adopt from one given by Stein [1962] for the problem of estimating the mean vector of a multivariate normal random variable. Let $\underset{\sim}{\lambda}$ denote a fixed vector of probabilities, i.e. $\underset{\sim}{\lambda} \in S_t$. $\underset{\sim}{\lambda}$ will be our choice

of "origin" within the parameter space; often we take $\underset{\sim}{\lambda} = \underset{\sim}{c}$, but in this development we let $\underset{\sim}{\lambda}$ denote a general choice of the origin. Now consider the triangle whose vertices are the three vectors $\underset{\sim}{\lambda}$, $\underset{\sim}{p}$ and $\hat{\underset{\sim}{p}}$. Let θ denote the angle between the vectors $\hat{\underset{\sim}{p}}-\underset{\sim}{p}$ and $\underset{\sim}{\lambda}-\underset{\sim}{p}$. Then, if $\underset{\sim}{p}$ is constrained to lie away from the boundary of S_t, it may be shown that

$$E(\cos^2(\theta)) = O(t^{-1}) \text{ as } t \to \infty. \tag{2.5}$$

Because θ is very nearly a right angle, there are points along the line connecting $\hat{\underset{\sim}{p}}$ and $\underset{\sim}{\lambda}$ that are closer to $\underset{\sim}{p}$ than is $\hat{\underset{\sim}{p}}$. This fact leads us to consider estimates of $\underset{\sim}{p}$ that are formed by shrinking $\hat{\underset{\sim}{p}}$ towards the origin, $\underset{\sim}{\lambda}$. Any point along the line connecting $\hat{\underset{\sim}{p}}$ and $\underset{\sim}{\lambda}$ may be represented as

$$w\hat{\underset{\sim}{p}} + (1-w)\underset{\sim}{\lambda} \text{ for } 0 \leq w \leq 1. \tag{2.6}$$

Expression (2.5) also coincides with the class of Bayes estimators of $\underset{\sim}{p}$, arrived at by using the family of Dirichlet prior distributions (e.g. see Good, 1965). If we write the Dirichlet prior density as

$$\Gamma(K) \prod_{i=1}^{t} p_i^{K\lambda_i - 1} / \Gamma(K\lambda_i), \tag{2.7}$$

then (2.6) is the form of the posterior mean of $\underset{\sim}{p}$, with

$$w = n/(n+K). \tag{2.8}$$

For a geometrical interpretation of K, see Fienberg and Holland [1972].

We denote estimators of $\underset{\sim}{p}$ that have the form (2.6) with w given by (2.8) as $\hat{\underset{\sim}{q}}(K,\underset{\sim}{\lambda})$. The risk function of $\hat{\underset{\sim}{q}}(K,\underset{\sim}{\lambda})$ is

$$R(\hat{\underset{\sim}{q}},\underset{\sim}{p}) = w(1-\|\underset{\sim}{p}\|^2) + (1-w)^2 n \|\underset{\sim}{p}-\underset{\sim}{\lambda}\|^2. \tag{2.9}$$

Various choices of K in (2.8) have appeared in the literature. For example, $K = \frac{1}{2}t$ corresponds to adding a pseudo count of $\frac{1}{2}t\lambda_i$ to each cell. When the λ_i are all equal this adds $\frac{1}{2}$ to each cell and we get the procedure suggested by Goodman. The estimator $\underset{\sim}{p}_M = \hat{\underset{\sim}{q}}(\sqrt{n}, \underset{\sim}{c})$ is the unique, constant risk, minimax estimator of $\underset{\sim}{p}$ (Steinhaus, 1957; Trybula, 1958). The risk of $\underset{\sim}{p}_M$ is

$$(\sqrt{n}/(\sqrt{n}+1))^2 \ (1 - (1/t)). \tag{2.10}$$

The value of w that minimizes the expected squared distance from $\hat{\underset{\sim}{q}}$ to $\underset{\sim}{p}$ is of the form (2.8) with

$$K = K(\underset{\sim}{p}, \underset{\sim}{\lambda}) = (1-||\underset{\sim}{p}||^2)/||\underset{\sim}{p}-\underset{\sim}{\lambda}||^2. \tag{2.11}$$

This value of K depends on the unknown $\underset{\sim}{p}$. In a previous paper (Fienberg and Holland, 1970), we suggested estimating this optimal value of K by its maximum likelihood estimate

$$\hat{K} = K(\hat{\underset{\sim}{p}}, \underset{\sim}{\lambda}) = \frac{n^2 - \sum_{i=1}^{t} X_i^2}{\sum_{i=1}^{t} X_i^2 - 2n \sum_{i=1}^{t} X_i\lambda_i + n^2 \sum_{i=1}^{t} \lambda_i^2}. \tag{2.12}$$

Our proposed class of estimators of $\underset{\sim}{p}$ is then given by

$$\underset{\sim}{p}^* = \hat{\underset{\sim}{q}}(\hat{K}, \underset{\sim}{\lambda}) = [n/(n+\hat{K})]\hat{\underset{\sim}{p}} + [\hat{K}/(n+\hat{K})]\underset{\sim}{\lambda} \tag{2.13}$$

where $\hat{K}$ is given by (2.12) and the probability vector $\underset{\sim}{\lambda}$ indexes the class as it ranges over S_t. We note that other estimates of $K(\underset{\sim}{p}, \underset{\sim}{\lambda})$ are possible, and these may lead to further improvements in the estimation of $\underset{\sim}{p}$.

3. Two-Stage Bayes Approach

In Section 2 we used a geometrical argument to motivate our estimator of the multinomial probability vector $\underset{\sim}{p}$. As we pointed out, the class of estimators given by (2.6) coincides with the class of Bayes estimators of $\underset{\sim}{p}$ where the prior density is Dirichlet. Following an argument of Good [1967], when $\underset{\sim}{X}$ is multinomial with parameters n and $\underset{\sim}{p}$, we let $\underset{\sim}{p}$ have a Type II Dirichlet prior distribution with parameters $(K,\underset{\sim}{\lambda})$ and density (2.7), and then let $(K,\underset{\sim}{\lambda})$ have a Type III prior distribution with density function $\varphi(K,\underset{\sim}{\lambda})$. Then the conditional expectation of $\underset{\sim}{p}$ given $\underset{\sim}{X}$ is

$$E(\underset{\sim}{p}|\underset{\sim}{X}) = w(\underset{\sim}{X})\hat{\underset{\sim}{p}} + (1-w(\underset{\sim}{X}))\underset{\sim}{\lambda}(\underset{\sim}{X}) \tag{3.1}$$

where

$$w(\underset{\sim}{X}) = n/(n+K(\underset{\sim}{X})), \tag{3.2}$$

$$K(\underset{\sim}{X}) = \frac{\int \frac{K}{n+K} H(\underset{\sim}{X},K,\underset{\sim}{\lambda})\varphi(K,\underset{\sim}{\lambda})\,dK d\underset{\sim}{\lambda}}{\int \frac{1}{n+K} H(\underset{\sim}{X},K,\underset{\sim}{\lambda})\varphi(K,\underset{\sim}{\lambda})\,dK d\underset{\sim}{\lambda}}, \tag{3.3}$$

$$\lambda_i(\underset{\sim}{X}) = \frac{\int \lambda_i \frac{K}{n+K} H(\underset{\sim}{X},K,\underset{\sim}{\lambda})\varphi(K,\underset{\sim}{\lambda})\,dK d\underset{\sim}{\lambda}}{\int \frac{K}{n+K} H(\underset{\sim}{X},K,\underset{\sim}{\lambda})\varphi(K,\underset{\sim}{\lambda})\,dK d\underset{\sim}{\lambda}}, \tag{3.4}$$

and $H(\underset{\sim}{X},K,\underset{\sim}{\lambda})$ is the Bayes factor

$$H(\underset{\sim}{X},K,\underset{\sim}{\lambda}) = \frac{\Gamma(K)}{\Gamma(n+K)} \prod_i \frac{\Gamma(X_i+K\lambda_i)}{\Gamma(K\lambda_i)}. \tag{3.5}$$

We can use this result to provide additional motivation for our estimators $\underset{\sim}{p}^*$. For suppose we can find a degenerate density $\varphi(K,\underset{\sim}{\lambda})$ concentrated on a line so that $\underset{\sim}{\lambda}$ is a constant, and such the $\underset{\sim}{K}(\underset{\sim}{X})$ is given by (2.12). Then $E(\underset{\sim}{p}|\underset{\sim}{X}) = \underset{\sim}{p}^*$ and $\underset{\sim}{p}^*$ would be a Bayes estimator using a two-stage prior. The two-stage Bayes approach also suggests that we let $\underset{\sim}{\lambda}$ depend on $\underset{\sim}{X}$ as well, and we consider this possibility in Section 6.

It is not likely that one can find a density $\varphi(K,\underset{\sim}{\lambda})$ such that $E(\underset{\sim}{p}|\underset{\sim}{X}) = \underset{\sim}{p}^*$, but this equality holds approximately for some φ. Hence, we refer to our estimator $\underset{\sim}{p}^*$ as pseudo-Bayes rather than Bayes. There is other work along similar lines by Novick, Lewis and Jackson [1971] and Leonard [1972]. These authors transform binomial and multinomial data using arc-sine and log-odds transformations. They assume multivariate normality for the transformed variables, and then use a general approach of Lindley [1971] for normally distributed random variables. Efron and Morris [1971,1972] have considered similar problems for the normal case in which the prior parameters are estimated from the marginal distribution of $\underset{\sim}{X}$. They refer to their estimators as empirical Bayes.

4. Asymptotic Results

The usual asymptotic approach to multinomial problems holds t fixed and lets the sample size n tend to infinity. We consider this approach in Section 4.2.

A different asymptotic approach lets both n and t tend to infinity, but at the same rate so that n/t remains constant. One reason for looking at this special type of asymptotics comes from practical considerations. Typically multinomial data comes in the form of a cross-classification of discrete variables. In many situations, there are a large number of variables which could be used to cross-classify each observation, and if all of these variables were used the data would be spread too thinly over the cells in the resulting multi-dimensional contingency table. Thus, if the investigator uses a subset of the variables to keep the average number of observations from becoming too small, he is in effect choosing t so that n/t is moderate. If t is large, then he is in the special type of asymptotic situation described in detail in Section 4.1.

4.1 Special asymptotics for sparse multinomials

The asymptotic set-up that describes a sparse multinomial distribution lets $t \to \infty$ and sets $n = \delta t$ where $\delta = n/t$ is a constant. The dimension, t, varies in this asymptotic set-up so that S_t, the parameter space, is also varying with t. Instead of having a single fixed probability vector, we must consider an infinite sequence of probability vectors whose dimensions increase without bound. This type of asymptotic set-up has been treated before, e.g. by Morris [1969]. We choose to simplify the structure of this situation by relating the elements of this sequence of probability vectors through the following device. Let $p(\cdot)$ denote a probability density function on [0,1].

For each value of t we let

$$p_i = (1/t)p(\tfrac{i-\frac{1}{2}}{t}) \quad i = 1,2,\ldots,t. \tag{4.1}$$

Strictly speaking p_i defined in (4.1) should depend explicitly on t (i.e. $p_{i,t}$) but this excessive notation will not be used here. Furthermore, the vector $\underset{\sim}{p}_t = (p_1,\ldots,p_t)$ defined from (4.1) is not necessarily an element of S_t since $\sum_{i=1}^{t} p_i$ need not be unity. However, if $p(\cdot)$ is sufficiently smooth (for example, if $p(\cdot)$ has a continuous second derivative) then standard results in numerical integration (Davis and Rabinowitz, 1967) show that

$$\sum_{i=1}^{t} p_i = \sum_{i=1}^{t} p(\tfrac{i-\frac{1}{2}}{t})(1/t) = \int_0^1 p(x)dx + o(t^{-1}) = 1 + o(t^{-1}). \tag{4.2}$$

This will be sufficient for our purposes.

Let $\lambda(\cdot)$ be a second probability density on [0,1] and set

$$\lambda_i = (1/t)\lambda(\tfrac{i-\frac{1}{2}}{t}) \quad i = 1,\ldots,t. \tag{4.3}$$

By assuming that both $p(\cdot)$ and $\lambda(\cdot)$ have continuous second derivatives it follows that for all $\alpha, \beta \geq 0$ we have

$$\sum_{i=1}^{t} p_i^{\alpha}\lambda_i^{\beta} = (1/t)^{\alpha+\beta-1}\int_0^1 p^{\alpha}(x)\lambda^{\beta}(x)dx + o(t^{-\alpha-\beta}) \tag{4.4}$$

Thus we can replace summations involving p_i and λ_i by integrals involving $p(\cdot)$ and $\lambda(\cdot)$. For example, the risk for $\hat{p}$ in (2.4) may be expressed as follows

$$R(\hat{\underset{\sim}{p}},\underset{\sim}{p}) = \sum_{i=1}^{t} p_i - \sum_{i=1}^{t} p_i^2 = \int_0^1 p(x)dx + o(t^{-1}) - (1/t)\int_0^1 p^2(x)dx + o(t^{-2}). \tag{4.5}$$

Henceforth, references to x, dx and the limits of integration will be omitted in all expressions using integrals, e.g.

$$R(\hat{\underset{\sim}{p}},\underset{\sim}{p}) = 1 - (1/t)\int p^2 + o(t^{-1}). \tag{4.6}$$

Equation (4.6) is prototypic of our expansions of risk functions. The expansion for the risk function (2.9) of $\hat{\underset{\sim}{q}}(K,\underset{\sim}{\lambda})$ depends on how K behaves as a function of t. If $K = \frac{1}{2}t$

$$R(\hat{\underset{\sim}{q}}(\tfrac{1}{2}t,\underset{\sim}{\lambda}),\underset{\sim}{p}) = w_0^2 + (1-w_0)^2 D - (1/t)w_0^2\int p^2 + o(t^{-1}) \tag{4.7}$$

where

$$w_0 = \delta/(\delta+\tfrac{1}{2}), \quad D = \delta\int(\lambda-p)^2. \tag{4.8}$$

The reader should note that as p and λ vary over all possible density functions on [0,1], D varies from 0 to ∞. The risk of the minimax estimator, $\underset{\sim}{p}_M$, given in (2.10) may be expressed as

$$1 - \frac{2}{\sqrt{n}} + \frac{1}{t}(\frac{3}{\delta} - 1) + o(t^{-1}). \tag{4.9}$$

For fixed $p(\cdot)$ and δ, the three expansions (4.6), (4.7) and (4.9) give the risk functions of three estimators of $\underset{\sim}{p}$ to order t^{-1}. We propose to compare estimators of $\underset{\sim}{p}$ on the basis of the leading term of these expansions.

Our main purpose in developing the asymptotics of sparse multinomial distributions is to approximate the risk function of $\underset{\sim}{p}^*$ given by (2.13). Holland and Sutherland [1971] have shown that, to order t^{-1}, the risk function of $\underset{\sim}{p}^*$ is

$$R(\underset{\sim}{p}^*,\underset{\sim}{p}) = S_0(D) + \frac{1}{n}S_1(D) + \frac{1}{t}\int p^2 S_2(D) + \frac{\delta}{t}S_3(D)\sigma^2(p-\lambda) + o(t^{-1}) \tag{4.10}$$

where

$$\begin{aligned} S_0(D) &= (D^2+3D+1)/(D+2)^2 \\ S_1(D) &= 2(D+1)/(D+2)^3 \\ S_2(D) &= -(D^4+6D^3+7D^2-6D-2)/(D+2)^4 \\ S_3(D) &= 4(D^2+3D-1)/(D+2)^4, \end{aligned} \tag{4.11}$$

$$\sigma^2(p-\lambda) = \int [p-\lambda-\mu(p-\lambda)]^2 p, \tag{4.12}$$

$$\mu(p-\lambda) = \int (p-\lambda)p. \tag{4.13}$$

and D is defined in (4.8).

For comparison we collect all of the leading terms of the risk function expansions given in this section in Table 4.1.

Table 4.1 goes here

It is evident that $S_0(D)$ in (4.11) satisfies the inequality

$$\tfrac{1}{4} \leq S_0(D) < 1 \tag{4.14}$$

for all $D \geq 0$, no matter what the choice of $\lambda(\cdot)$. Hence this first order analysis shows that $\underset{\sim}{p}^*$ has a risk function whose leading term is uniformly smaller than that of $\hat{\underset{\sim}{p}}$ for all $p(\cdot)$. This implies that, for all fixed $p(\cdot)$ and δ, if t is large enough the risk of $\underset{\sim}{p}^*$ is less than that of $\hat{\underset{\sim}{p}}$. Thus the maximum likelihood estimator, $\hat{p}$, is "asymptotically inadmissible" in the framework of our special asymptotics for sparse multinomials. This "asymptotic inadmissibility" is not in conflict with the "regular admissibility" of $\hat{p}$, as proved by Johnson [1971] for fixed dimensionality, but merely reflects the fact that $\hat{\underset{\sim}{p}}$ gains its regular admissibility only from its behavior on the boundary of the parameter space. The special asymptotic analysis for sparse multinomials emphasizes the interior of the parameter space, and, in effect, ignores the boundary.

Figure 4.1 graphs the leading terms in Table 4.1 for $\delta = 5$, $n = 100$. Here $\underset{\sim}{p}^*$, our pseudo-Bayes estimator, has smaller risk than $\hat{\underset{\sim}{q}}(\tfrac{1}{2}t,\underset{\sim}{\lambda})$, the estimator formed by adding $\tfrac{1}{2}$ to each cell, when t is large, for any $p(\cdot)$.

Figure 4.1 goes here

4.2 Standard asymptotics for multinomials

In this section we regard t as fixed and study $\underset{\sim}{p}^*$ and $\hat{\underset{\sim}{p}}$ as $n \to \infty$. In order to allow for the situation described in Section 6 we modify our notation for $\underset{\sim}{\lambda}$ somewhat. We shall let $\underset{\sim}{\lambda}$ depend on $\underset{\sim}{X}$ and denote this by $\hat{\underset{\sim}{\lambda}} = \underset{\sim}{\lambda}(\underset{\sim}{X})$. Furthermore, we assume that there is a function of $\underset{\sim}{p}$, $\underset{\sim}{\lambda}^* = \underset{\sim}{\lambda}^*(\underset{\sim}{p})$ such that $\sqrt{n}(\hat{\underset{\sim}{\lambda}}-\underset{\sim}{\lambda}^*)$

has an asymptotic, possibly degenerate, multivariate Normal distribution with mean zero as $n \to \infty$. In Section 6 the choice of $\hat{\underset{\sim}{\lambda}}$ that receives the most attention is $\hat{\lambda}_{ij} = X_{i+}X_{+j}/n^2$ with $\lambda^*_{ij}(\underset{\sim}{p}) = p_{i+}p_{+j}$. It is well-known in this case that $\sqrt{n}(\hat{\underset{\sim}{\lambda}}-\underset{\sim}{\lambda}^*)$ does have an asymptotic multivariate Normal distribution. If $\hat{\underset{\sim}{\lambda}}$ is a constant (i.e. non-random) then $\sqrt{n}(\hat{\underset{\sim}{\lambda}}-\underset{\sim}{\lambda}^*) \equiv 0$ which we shall interpret as degenerate asymptotic Normality. Thus our present notation will not conflict with our previous assumption that $\underset{\sim}{\lambda}$ is non-random.

Let $\hat{w}_n = n/(n+\hat{K})$ where $\hat{K}$ is given by (2.12). Since $\hat{\underset{\sim}{\lambda}}$ converges in probability to $\underset{\sim}{\lambda}^*$ we have the following lemma, given here without proof.

Lemma 4.1: If $\underset{\sim}{p} \neq \underset{\sim}{\lambda}^*(\underset{\sim}{p})$ then $\hat{K}$ converges in probability to $(1-||\underset{\sim}{p}||^2)/||\underset{\sim}{\lambda}^*-\underset{\sim}{p}||^2$ and consequently $1 - \hat{w}_n = 0_p(n^{-1})$.

Now let $\underset{\sim}{U}_n = \sqrt{n}(\hat{\underset{\sim}{\lambda}}-\underset{\sim}{\lambda}^*)$, $\underset{\sim}{V}_n = \sqrt{n}(\hat{\underset{\sim}{p}}-\underset{\sim}{p})$ and assume that $(\underset{\sim}{U}_n,\underset{\sim}{V}_n)$ converges in distribution to $(\underset{\sim}{U},\underset{\sim}{V})$ which has a possibly degenerate multivariate Normal distribution with zero mean. The next lemma compliments Lemma 4.1 when $\underset{\sim}{p}=\underset{\sim}{\lambda}^*$.

Lemma 4.2: If $\underset{\sim}{p} = \underset{\sim}{\lambda}^*(\underset{\sim}{p})$ then $\hat{w}_n$ converges in distribution to the random variable

$$w = \frac{||\underset{\sim}{U} - \underset{\sim}{V}||^2}{1 - ||\underset{\sim}{p}||^2 + ||\underset{\sim}{U} - \underset{\sim}{V}||^2} . \qquad (4.15)$$

These lemmas may be used to show that if $\underset{\sim}{p} \neq \underset{\sim}{\lambda}^*$ then $\underset{\sim}{p}^*$ is asymptotically equivalent to $\hat{\underset{\sim}{p}}$ and therefore that $\underset{\sim}{p}^*$ is a consistent estimate of $\underset{\sim}{p}$ as $n \to \infty$. The easy proof of this result is given in Fienberg and Holland [1970].

Theorem 4.1: (a) If $\underset{\sim}{p} \neq \underset{\sim}{\lambda}^*$, then $\sqrt{n}(\underset{\sim}{p}^*-\hat{\underset{\sim}{p}}) = 0_p(n^{-1})$.

(b) If $\underset{\sim}{p} = \underset{\sim}{\lambda}^*$, then $\sqrt{n}(\underset{\sim}{p}^*-\hat{\underset{\sim}{p}})$ converges in distribution to $(1-w)(\underset{\sim}{U}-\underset{\sim}{V})$ where w is given by (4.15).

Corollary: $\underset{\sim}{p}^*$ is a consistent estimator of $\underset{\sim}{p}$ as $n \to \infty$.

From part (a) of Theorem 4.1 we see that if $\underset{\sim}{p} \neq \underset{\sim}{\lambda}^*$ then $\underset{\sim}{p}^*$ is asymptotically equivalent to $\hat{\underset{\sim}{p}}$ in the sense that they have the same asymptotic distributions. Hence the ratio of the risk of $\underset{\sim}{p}^*$ to that of $\hat{\underset{\sim}{p}}$ approaches unity as $n \to \infty$. The situation is more complicated when $\underset{\sim}{p} = \underset{\sim}{\lambda}^*$. From part (b) of the theorem it follows that if $\underset{\sim}{p} = \underset{\sim}{\lambda}^*$ then the ratio of the risks converges to the value of

$$(1 - \|\underset{\sim}{p}\|^2)^{-1} \, E\|w\underset{\sim}{V} + (1-w)\underset{\sim}{U}\|^2. \tag{4.16}$$

When $\hat{\underset{\sim}{\lambda}} \equiv \underset{\sim}{\lambda}^* = \underset{\sim}{p} = \underset{\sim}{c}$, $\underset{\sim}{U} \equiv 0$ and (4.16) becomes

$$E\left[\left(\frac{t\|\underset{\sim}{V}\|^2}{t\|\underset{\sim}{V}\|^2 + t - 1} \right)^2 \frac{t\|\underset{\sim}{V}\|^2}{t-1} \right]. \tag{4.17}$$

Using the asymptotic (as $n \to \infty$) multivariate Normality of $\underset{\sim}{V}$, and a standard orthogonality argument, we can expand (4.17) directly to order t^{-1} yielding

$$\tfrac{1}{4} + (3/8)t + o(t^{-1}). \tag{4.18}$$

Professor R.R. Bahadur (personal communication) has pointed out that both the upper and lower bounds for (4.17) are given by

$$\frac{1}{4} + \frac{3}{8(t-1)} - \frac{1}{2(t-1)^2} + \theta \frac{9}{4(t-1)^2}\left[1 + \frac{2}{t-1}\right], \tag{4.19}$$

where $0 < \theta < 1$. To order t^{-1} (4.19) is consistent with (4.18).

If we divide the expansion (4.10) from Section 4.1 by $R(\hat{\underset{\sim}{p}}, \underset{\sim}{p})$ and set D equal to zero, λ equal to 1 and let δ go to infinity, then we also obtain (4.18). From (4.18) we see that when $\underset{\sim}{p} = \underset{\sim}{\lambda} = \underset{\sim}{c}$ the asymptotic risk ratio converges to .25 from above as t gets large. The actual rate of this convergence is indicated in Section 5.3.

5. Some Small Sample Results

As noted in Section 4, the risk of $\underset{\sim}{p}^*$ does not have a convenient algebraic closed-form due to the dependence of $\hat{K}$ on $\underset{\sim}{X}$. We have already noted that as $n \to \infty$ and $t \to \infty$ with n/t fixed, $\underset{\sim}{p}^*$ provides an almost uniform improvement over $\hat{\underset{\sim}{p}}$. Two questions remain:

(a) Does this large sample improvement carry over into small samples in any way?

(b) How large do n and t have to be, before the special asymptotics are meaningful?

Here we give some exact comparisons of the risk functions of $\hat{\underset{\sim}{p}}$ and $\underset{\sim}{p}^*$ in an attempt to answer these questions. In all our examples we take $\underset{\sim}{\lambda} = \underset{\sim}{c}$. The exact values of $R(\underset{\sim}{p}^*, \underset{\sim}{p})$ were evaluated numerically by high-speed computer.

5.1 Binomial Case

Figure 5.1 shows the risk of $\underset{\sim}{p}^*$, $\hat{\underset{\sim}{p}}$ and the minimax estimator, $\underset{\sim}{p}_M$, for $t = 2$ (the binomial case) and $n = 15$. Here $\underset{\sim}{p}^*$ has smaller risk than $\hat{\underset{\sim}{p}}$ for p_1 (and thus p_2) between 0.33 and 0.67 and has larger risk elsewhere. The minimax estimator, $\underset{\sim}{p}_M$, is superior to $\hat{\underset{\sim}{p}}$ between 0.2 and 0.8, but is dominated by $\underset{\sim}{p}^*$ between 0.42 and 0.58 and near the boundary where p_1 is near 0 or 1.

The behavior of $\underset{\sim}{p}_M$ and $\underset{\sim}{p}^*$ near the boundary deserves further comment. $\hat{\underset{\sim}{p}}$ dominates $\underset{\sim}{p}^*$ near the boundary, but both of their risks tend to 0 as p_1 approaches zero or one. Thus, the behavior of $\underset{\sim}{p}^*$ near the boundary is at least satisfactory. On the other hand, the ratio of the risk of $\underset{\sim}{p}_M$ to either that of $\underset{\sim}{p}^*$ or that of $\hat{\underset{\sim}{p}}$ tends to ∞ as we near the boundary, and for some purposes this may be unacceptable. This behavior of $\underset{\sim}{p}_M$ near the boundary is typical of all true Bayes estimators of the form (2.6) where w is constant. We do not consider $\underset{\sim}{p}_M$ in the subsequent examples.

Finally we note that, as $n \to \infty$ with $t = 2$ fixed, the "ears" of the risk function of $\underset{\sim}{p}^*$ rising above the risk function of $\hat{\underset{\sim}{p}}$ move toward each other and the difference between the functions disappears, except at $\underset{\sim}{p} = (\frac{1}{2},\frac{1}{2})$ where the risk of $\underset{\sim}{p}^*$ is strictly less than the risk of $\hat{\underset{\sim}{p}}$.

Figure 5.1 goes here

5.2 Trinomial Case

A three-dimensional picture is required to display the risk functions of $\underset{\sim}{p}^*$ and $\hat{\underset{\sim}{p}}$ for $t = 3$. In Fienberg and Holland [1970], rather than looking at these functions as they sit over the two-dimensional probability simplex, S_3, we presented sections along the two lines in S_3 defined by $[p_1, (1-p_1)/2, (1-p_1)/2]$ and $[p_1, 1-p_1, 0]$. Here we consider the ratio of the risks of $\underset{\sim}{p}^*$ and $\hat{\underset{\sim}{p}}$, and plot contours of constant risk ratio, i.e. contours for which

$$\rho(\underset{\sim}{p}) = \frac{R(\underset{\sim}{p}^*,\underset{\sim}{p})}{R(\hat{\underset{\sim}{p}},\underset{\sim}{p})} \qquad (5.1)$$

is constant. When this ratio is less than 1, $\underset{\sim}{p}^*$ is superior to $\hat{\underset{\sim}{p}}$.

Figures 5.2 and 5.3 give contours of $\rho(\underset{\sim}{p})$ in S_3 for $n = 15$ and $n = 30$ respectively. As before the greatest improvement of $\underset{\sim}{p}^*$ over $\hat{\underset{\sim}{p}}$ occurs at the point $\underset{\sim}{p} = \underset{\sim}{\lambda} = (1/3,1/3,1/3)$, where the value of $\rho(\underset{\sim}{p})$ equals 0.42 and 0.39 for $n = 15$ and $n = 30$. Over a large section of the parameter space $\underset{\sim}{p}^*$ has smaller risk than $\hat{\underset{\sim}{p}}$. Although $\hat{\underset{\sim}{p}}$ has smaller risk near the vertices of S_3, the value of $\rho(\underset{\sim}{p})$ near the vertices reaches maxima of 1.25 and 1.17 for $n = 15$ and $n = 30$, and then decreases. Interestingly, $\underset{\sim}{p}^*$ dominates $\hat{\underset{\sim}{p}}$ for sizeable segments of the one-dimensional boundaries.

Figures 5.2 and 5.3 go here

5.3 Keeping δ fixed

To check on the rate of approach to the special asymptotic behavior of $\rho(\underset{\sim}{p})$ described in Section 4.1, we focus on $\underset{\sim}{p} = \underset{\sim}{\lambda} = \underset{\sim}{c}$ and examine values of $\rho(\underset{\sim}{c})$

as t increases for various fixed values of δ. Figure 5.4 displays exact computations of $\rho(\underset{\sim}{c})$ for $\delta = 1,2,3,5,10$. By the time t reaches 13, $\rho(\underset{\sim}{c})$ is less than 0.30 for all integral values of δ, and we have come close to the asymptotic value of 0.25. The larger the value of δ, the faster $\rho(\underset{\sim}{c})$ approaches the asymptote for increasing t. We conjecture that similar rates of approach to asymptotic values occur for other values of $\underset{\sim}{p}$.

Figure 5.4 goes here

6. Random $\underset{\sim}{\lambda}$'s

For regular multinomial data it was natural for us to choose $\underset{\sim}{\lambda} = \underset{\sim}{c}$ by symmetry considerations. When the multinomial represents cross-classified data such a choice is not the most natural, and it is quite reasonable for us to find a $\underset{\sim}{\lambda}$ which either (a) reflects some prior knowledge about the cross-classification structure, or (b) represents some special cross-classification structure which can serve as a "null" model for $\underset{\sim}{p}$, often characterized by symmetry considerations.

In Section 3 we saw that a two-stage Bayesian argument leads to estimators of the form (2.6) with w given by (2.8) and with both K and $\underset{\sim}{\lambda}$ as functions of $\underset{\sim}{X}$. If our choice of such a random $\underset{\sim}{\lambda}$ is judicious, $\underset{\sim}{\lambda}$ may remain close to $\underset{\sim}{p}$ for a large portion of S_t, and particularly for those points away from the boundary where $\hat{\underset{\sim}{p}}$ gains its admissibility.

In this section we consider an rxc cross-classification (i.e. a contingency table with r rows and c columns) so that t = rc. Although probabilities and other quantities for two-way contingency tables are normally doubly subscripted, they can still be strung out in vector form, so we need not change the notation established in earlier sections, except in that we replace single subscripts by double ones.

We continue to work with an estimator of the form

$$\underset{\sim}{p}^* = [n/(n+\hat{K})]\hat{\underset{\sim}{p}} + [\hat{K}/(n+\hat{K})]\underset{\sim}{\lambda} \tag{6.1}$$

where

$$\hat{K} = K(\hat{\underset{\sim}{p}},\underset{\sim}{\lambda}) = \frac{n^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}^2}{\sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}^2 - 2n\sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}\lambda_{ij} - n^2\sum_{i=1}^{r}\sum_{j=1}^{c}\lambda_{ij}^2}. \tag{6.2}$$

First we decompose $\underset{\sim}{\lambda}$ into three components: (a) the row totals λ_{i+}, (b) the column totals λ_{+j}, and (c) the cross-product ratios:

$$\gamma_{ij} = \lambda_{ij}\lambda_{i+1,j+1}/\lambda_{i+1,j}\lambda_{i,j+1} \tag{6.3}$$

for $i = 1,2,\ldots,r-1$; $j = 1,2,\ldots,c-1$. A natural choice for the row and column totals of $\underset{\sim}{\lambda}$ is

$$\begin{aligned} \lambda_{i+} &= X_{i+}/n \quad i = 1,2,\ldots,r, \\ \lambda_{+j} &= X_{+j}/n \quad j = 1,2,\ldots,c, \end{aligned} \tag{6.4}$$

because of the interest in margin preservation (see Mosteller, 1968). Choosing $\gamma_{ij} = 1$ for all i and j corresponds to the independence of the variables corresponding to rows and columns, and leads to the usual "expected values" for computing chi-square, i.e.

$$\lambda^*_{ij} = X_{i+}X_{+j}/n^2. \tag{6.5}$$

Other values of $\underset{\sim}{\gamma}$ may be suitable and we can combine them with the margins (6.4) via the Deming-Stephan iterative proportional fitting procedure to yield $\underset{\sim}{\lambda}$ (see Mosteller, 1968, and Fienberg, 1970).

We now report on some exact comparisons of $\hat{\underset{\sim}{p}}$ and $\underset{\sim}{p}^*$, where $\underset{\sim}{p}^*$ is the estimate formed by using the random λ^* given by (6.5) in the case of a 2x2 table with $n = 20$. The parameter space S_4 is now a tetrahedron, and we choose to view it as being composed of surfaces on which the cross-product ratio, $\alpha = p_{11}p_{22}/p_{12}p_{21}$, is a constant (Fienberg and Gilbert, 1970). Due to symmetry considerations it suffices to look only at values of $\alpha \geq 1$. Each of the surfaces of constant α can be mapped one-to-one onto a square (see Fienberg, 1970), for which the coordinates of any point are given by the marginal totals of the probabilities in the 2x2 table.

Figures 6.1, 6.2 and 6.3 go about here

Figures 6.1, 6.2 and 6.3 give contours of constant risk ratio,

$$\tilde{\rho}(\underset{\sim}{p}) = \frac{R(\tilde{\underset{\sim}{p}}^*, \underset{\sim}{p})}{R(\hat{\underset{\sim}{p}}, \underset{\sim}{p})} , \qquad (6.6)$$

for α = 1, 3 and 5 respectively. Recall that $\tilde{\underset{\sim}{p}}^*$ has smaller risk than $\hat{\underset{\sim}{p}}$ when $\tilde{\rho}(\underset{\sim}{p})$ is less than unity. For α = 1 (Figure 6.1) we see that $\tilde{\rho}(\underset{\sim}{p})$ has a minimum value of .76 when $p_{1+} = p_{+1} = 0.5$, and nowhere is it greater than one. For α = 3 (Figure 6.2), $\tilde{\rho}(\underset{\sim}{p})$ is still less than unity everywhere, but the minimum value is approximately .92 at four symmetrically placed points, corresponding roughly to $p_{1+} = 0.50$ and $p_{+1} = 0.15$. For α = 5 (Figure 6.3) $\hat{\underset{\sim}{p}}$ begins to show some superiority over $\tilde{\underset{\sim}{p}}^*$ near the center of the surface, although there still remain values of $\underset{\sim}{p}$ for which $\tilde{\rho}(\underset{\sim}{p}) < 1$ and $\tilde{\underset{\sim}{p}}^*$ is superior to $\hat{\underset{\sim}{p}}$.

For other values of $\alpha > 5$ (not shown here) it appears that $\tilde{\rho}(\underset{\sim}{p})$ has a maximum value of slightly more than 1.22, and, for surfaces corresponding to $\alpha > 20$, the maximum value of $\tilde{\rho}(\underset{\sim}{p})$ appears to be decreasing, tending to 1 as $\alpha \to \infty$.

For a sizeable amount of the tetrahedron surrounding the <u>surface of independence</u> (α=1) our weighted biased estimator, $\tilde{\underset{\sim}{p}}^*$, has smaller risk than $\hat{\underset{\sim}{p}}$, the unrestricted maximum likelihood estimator. As the dimensions of the table increase so that rc/n remains constant, we conjecture that the hypervolume of region in which $\tilde{\underset{\sim}{p}}^*$ is superior to $\hat{\underset{\sim}{p}}$ to increase relative to the hypervolume of the entire simplex.

References

[1] Davis, P.J. and Rabinowitz, P., Numerical Integration, Waltham, Mass.: Blaisdell Publishing Co., (1967), 15-17 and 108-112.

[2] Efron, B. and Morris, C., "Limiting the risk of Bayes and empirical Bayes estimators--Part I: the Bayes case," Journal of the American Statistical Association, 66 (December 1971), 807-815.

[3] Efron, B. and Morris, C., "Limiting the risk of Bayes and empirical Bayes estimators--Part II: the empirical Bayes case," Journal of the American Statistical Association, 67 (March 1972), 130-139.

[4] Fienberg, S.E., "An iterative procedure for estimation in contingency tables," Annals of Mathematical Statistics, 41 (June 1970), 907-917.

[5] Fienberg, S.E. and Gilbert, J.P., "Geometry of a 2x2 contingency table," Journal of the American Statistical Association, 65 (June 1970), 694-701.

[6] Fienberg, S.E. and Holland, P.W., "Methods for eliminating zero counts in contingency tables." In Random Counts on Models and Structures (G.P. Patil, ed.), University Park, Pa.: Pennsylvania State University Press, 233-260.

[7] Fienberg, S.E. and Holland, P.W., "On the choice of flattening constants for estimating multinomial probabilities," Journal of Multivariate Analysis, 2 (March 1972), 127-134.

[8] Freeman, M.F. and Tukey, J.W., "Transformations related to the angular and the square root," Annals of Mathematical Statistics, 21 (December 1950), 607-611.

[9] Gart, J.J. and Zweifel, J.R., "On the bias of various estimators of the logit and its variance with application to quantal bioassay," Biometrika, 54 (June 1967), 181-187.

[10] Good, I.J., The Estimation of Probabilities, Cambridge, Mass.: Massachusetts Institute of Technology Press, (1965).

[11] Good, I.J., "A Bayesian significance test for multinomial distributions (with discussion)," Journal of the Royal Statistical Society, Series B, 29, No. 3 (1967), 399-431.

[12] Goodman, L.A., "Interactions in multidimensional contingency tables," Annals of Mathematical Statistics, 35 (June 1964), 632-646.

[13] Goodman, L.A., "The multivariate analysis of qualitative data, interactions among multiple cross-classifications," Journal of the American Statistical Association, 65 (March 1970), 226-256.

[14] Goodman, L.A., "The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications," Technometrics, 13 (February 1971), 33-62.

[15] Grizzle, J.E., Starmer, C.F., and Koch, G.G., "Analysis of categorical data by linear models," Biometrics, 25 (September 1969), 489-504.

[16] Holland, P.W. and Sutherland, M.R., "The risk of the Fienberg-Holland estimator," Harvard University, Department of Statistics Memo NS-160, (1971).

[17] Johnson, B. McK., "On the admissible estimators for certain fixed sample binomial problems," Annals of Mathematical Statistics, 42 (October 1971), 1579-1587.

[18] Johnson, W.D. and Koch, G.G., "Analysis of qualitative data: linear functions," Health Services Review, -- (Winter 1968), 358-369.

[19] Leonard, T., "Bayesian methods for multinomial data," ACT Technical Bulletin No. 4, Iowa City, Iowa: The American College Testing Program, (1972).

[20] Lindley, D.V., "The estimation of many parameters." In Foundations of Statistical Inference (V.P. Godambe and D.A. Sprott, eds.), Toronto: Holt, Rinehart, and Winston, (1971), 435-455.

[21] Morris, C., "Central limit theorems for multinomial sums," Rand Corporation Technical Report RM-6026-PR, (1969).

[22] Mosteller, F., "Association and estimation in contingency tables," Journal of the American Statistical Association, 63 (March 1968), 1-28.

[23] Novick, M.R., Lewis, C., and Jackson, P.H., "Estimation of proportions in m groups," ACT Technical Bulletin No. 1, Iowa City, Iowa: The American College Testing Program (1971).

[24] Stein, C., "Confidence sets for the mean of a multivariate normal distribution (with discussion)," Journal of the Royal Statistical Society, Series B, 24, No. 2 (1962), 265-296.

[25] Steinhaus, H., "The problem of estimation," Annals of Mathematical Statistics, 28 (September 1957), 633-648.

[26] Trybula, S., "Some problems of simultaneous minimax estimation," Annals of Mathematical Statistics, 29 (March 1958), 245-253.

Table 4.1

Estimator	Leading term
$\hat{\underset{\sim}{p}}$	1
$p_M = \hat{\underset{\sim}{q}}(\sqrt{n}, \underset{\sim}{c})$	$1 - \frac{2}{\sqrt{n}}$
$\hat{\underset{\sim}{q}}(\frac{1}{2}t, \underset{\sim}{\lambda})$	$w_0^2 + (1-w_0)^2 D$
$\underset{\sim}{p}^*$	$(D^2+3D+1)/(D+2)^2$

Leading term in expansion of risk function for four estimators of $\underset{\sim}{p}$.

Legends for Figures

Fig 4.1 Leading terms of risk functions (δ=5, n=100) for four estimators of $\underset{\sim}{p}$. 1=$\hat{\underset{\sim}{p}}$, 2=$\hat{\underset{\sim}{q}}(\frac{1}{2}t,\underset{\sim}{\lambda})$, 3=$\hat{\underset{\sim}{q}}(\sqrt{n},\underset{\sim}{c})=\underset{\sim}{p}_M$, 4=$\underset{\sim}{p}^*$.

Fig 5.1 Risk of $\hat{\underset{\sim}{p}}$, $\underset{\sim}{p}_M$ and $\underset{\sim}{p}^*$ for t=2 and n=15.

Fig 5.2 Contours of constant risk ratio ($\underset{\sim}{p}^*$ over $\hat{\underset{\sim}{p}}$) for t=3 and n=15.

Fig 5.3 Contours of constant risk ratio ($\underset{\sim}{p}^*$ over $\hat{\underset{\sim}{p}}$) for t=3 and n=30.

Fig 5.4 Risk ratio ($\underset{\sim}{p}^*$ over $\hat{\underset{\sim}{p}}$) at the center of the simplex for various values of t and δ.

Fig 6.1 Countours of constant risk ratio ($\tilde{\underset{\sim}{p}}^*$ over $\hat{\underset{\sim}{p}}$) in the 2x2 table. n=20, α=1.

Fig 6.2 Contours of constant risk ratio ($\tilde{\underset{\sim}{p}}^*$ over $\hat{\underset{\sim}{p}}$) in the 2x2 table. n=20, α=3.

Fig 6.3 Contours of constant risk ratio ($\tilde{\underset{\sim}{p}}^*$ over $\hat{\underset{\sim}{p}}$) in the 2x2 table. n=0, α=5.

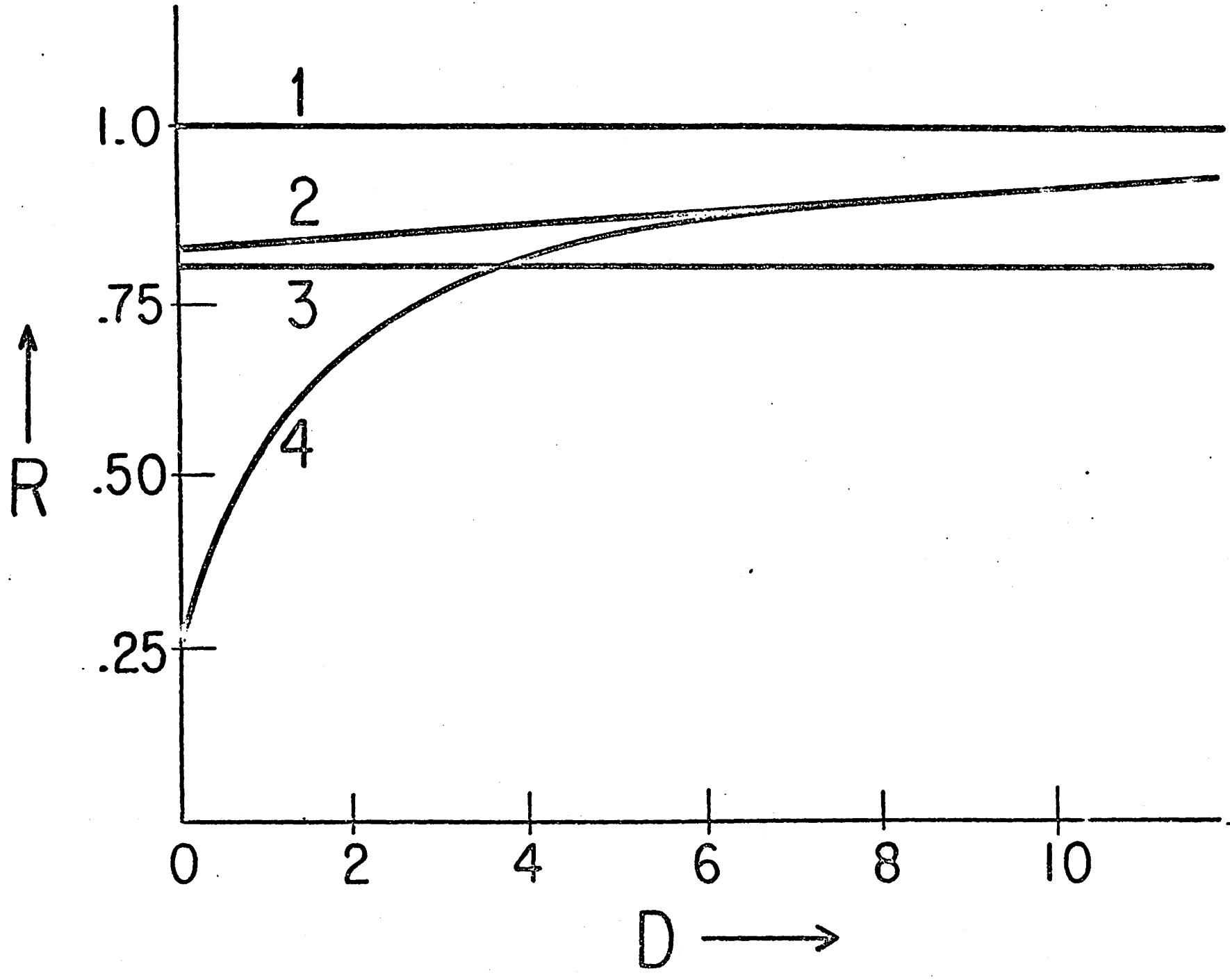

Fig 4.1

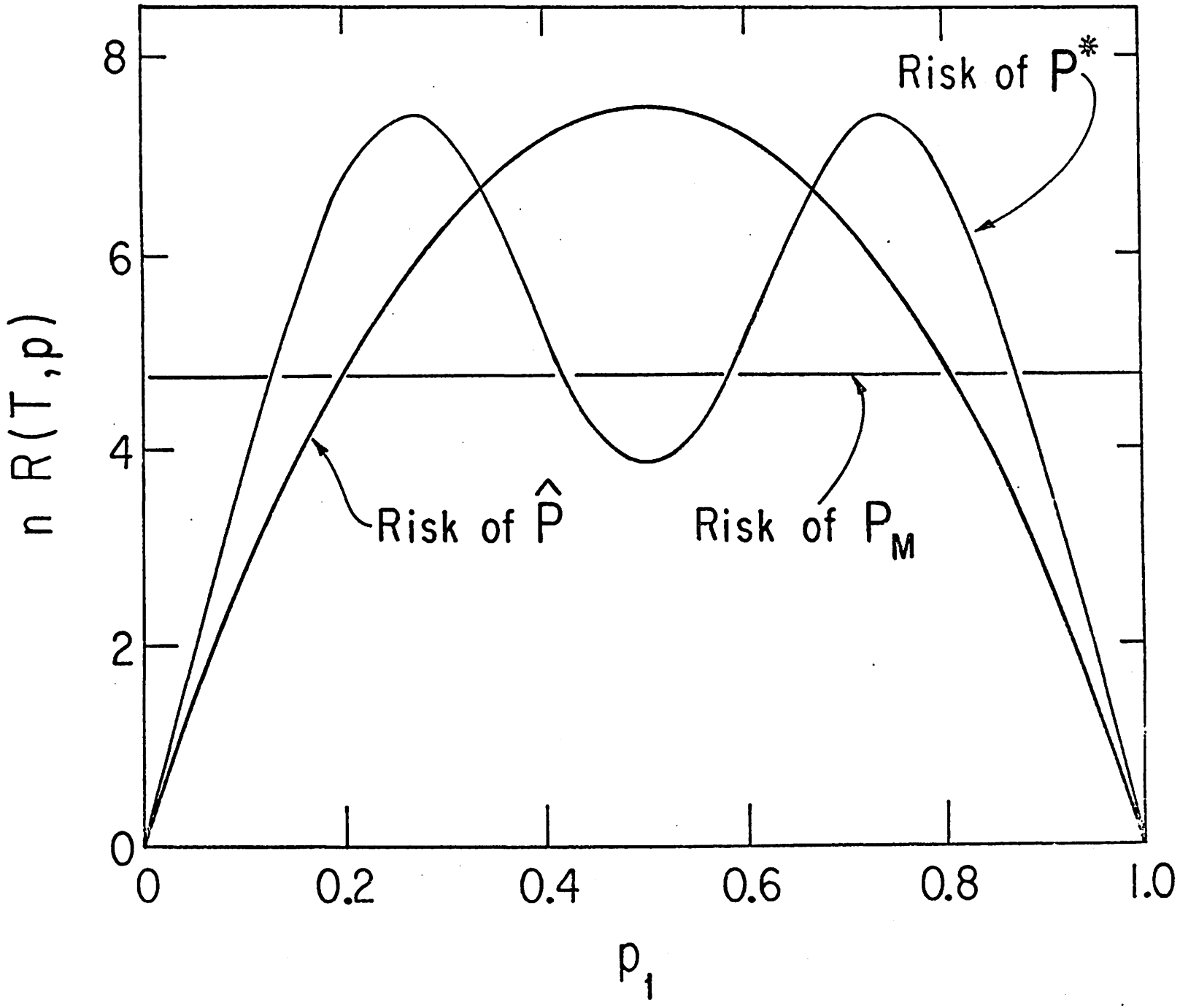

Fig 5.1

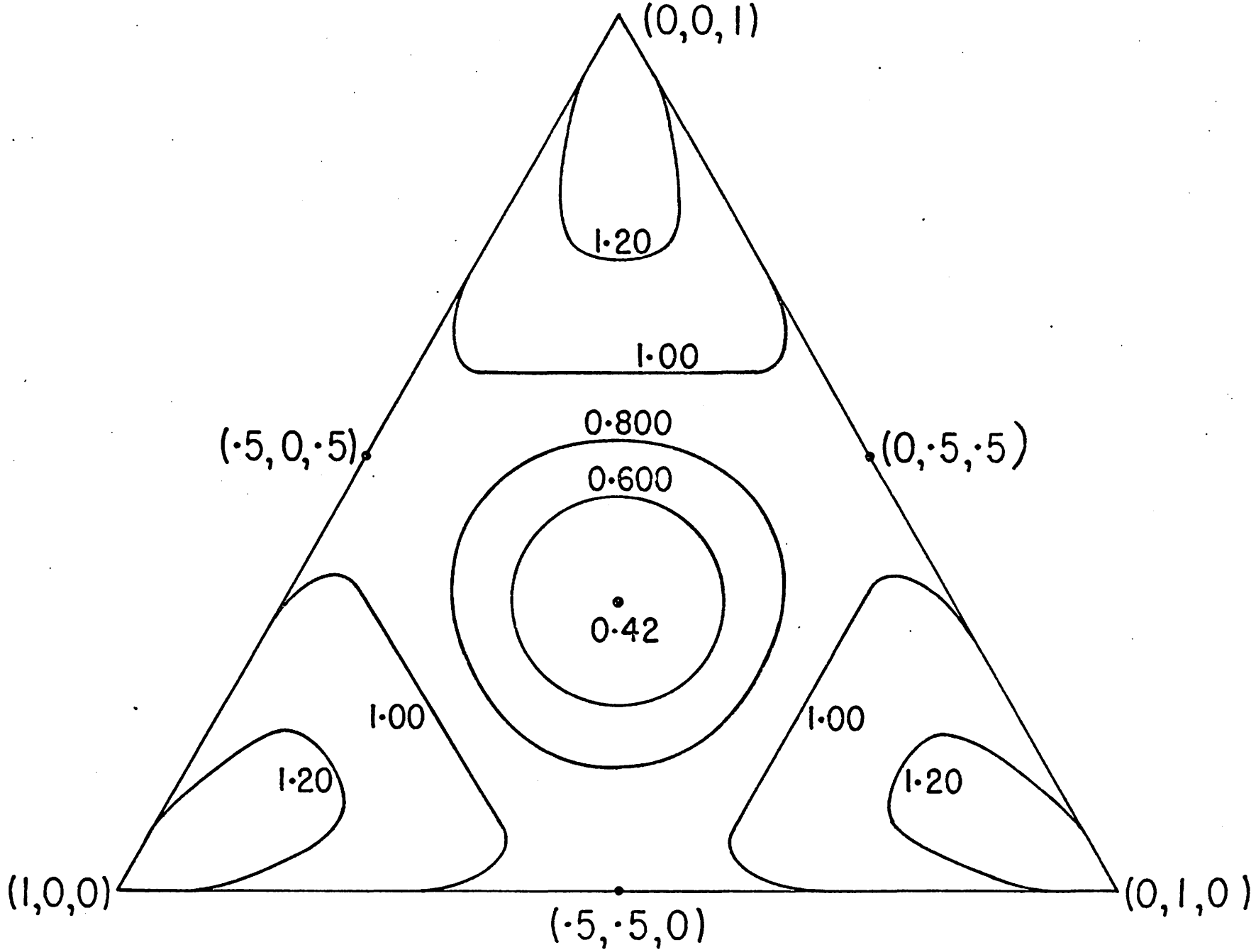

Fig 5.2

Fig 5.3

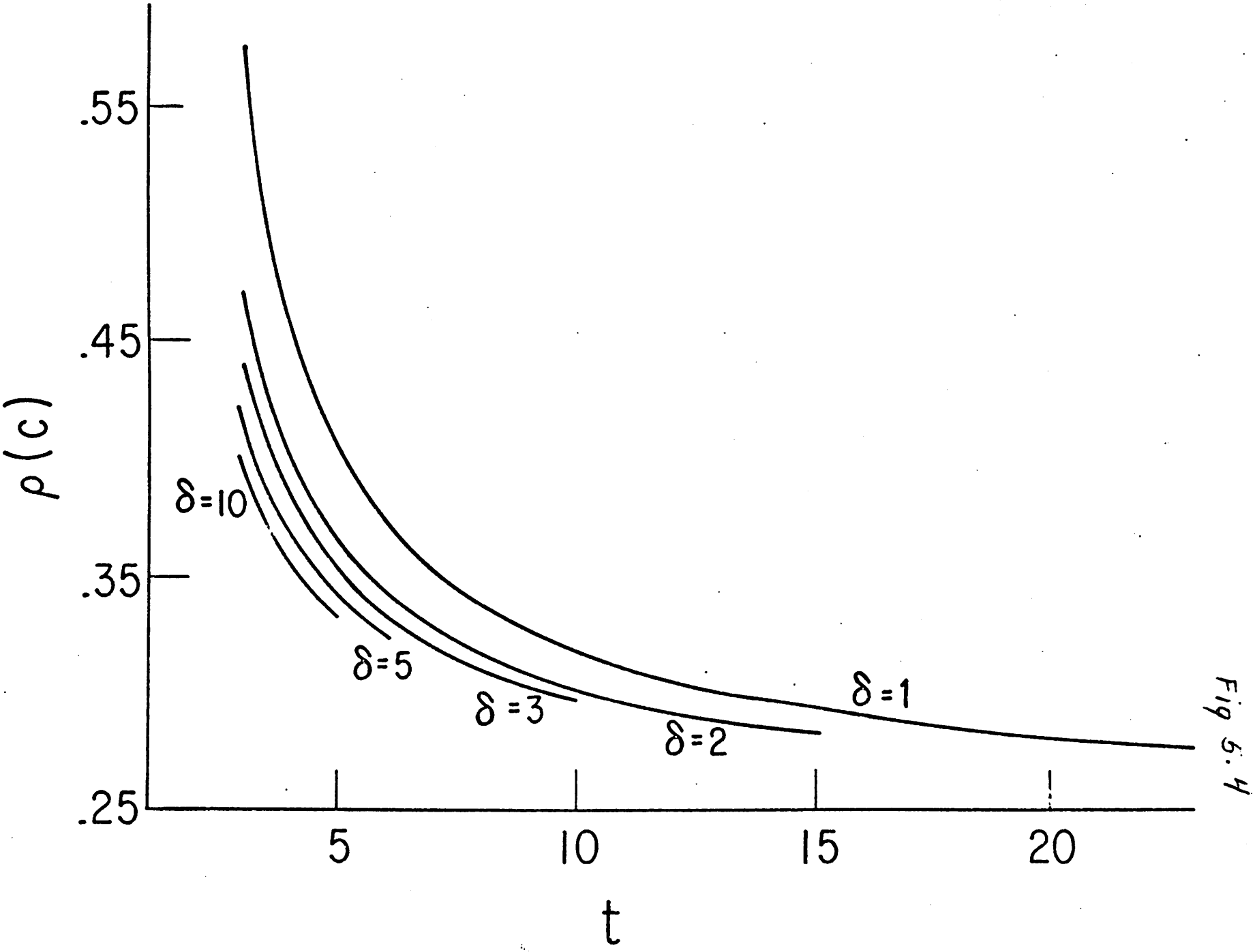

Fig 5.4

Fig 6.1

Fig 6.2

Fig 6.3